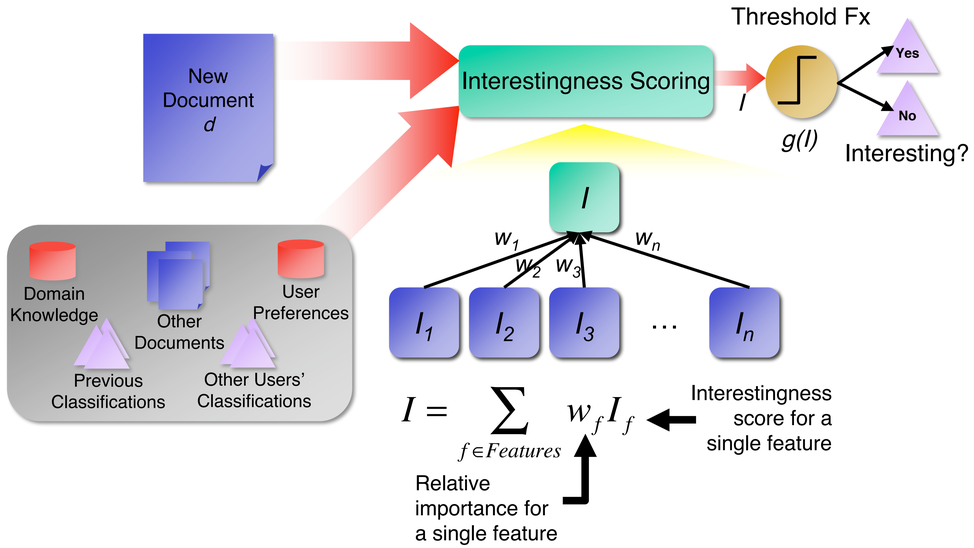
Measuring and Quantifying Interestingness of Documents

Raymond K. Pon, UC Los Angeles

David Buttler and Terence Critchlow, Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

Search engines, such as Google, assign scores to documents, such as news articles, based on their relevancy to a particular query or topic. However, not all relevant documents may be interesting to the user, particularly if the document is old or yields very little new information. Furthermore, relevancy scores do not take into account the uniqueness and the impact to a domain of interest by the document. To remedy this situation, we present a general framework for defining and measuring the "interestingness" of documents taking into account the above document features along with the event the document describes, document source analysis, user-feedback, and collaborative scoring. Applications include intelligent news aggregators, automated event detection and monitoring, and personalized web portals.

Document Classification



Training with Gradient Descent

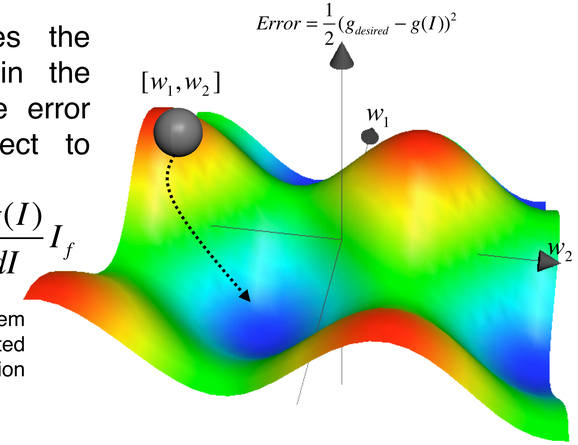
Gradient descent moves the weight assignment w_f in the opposite direction of the error change rate with respect to weight assignments

$$\Delta w_f = \eta(g_{desired} - g(I))\frac{dg(I)}{dI} I_f$$

Learning rate (step value)

Correct classification

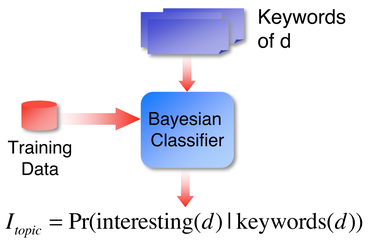
System generated classification



Moving weight assignments for two features to a region with lower error.

What Makes a Document Interesting?

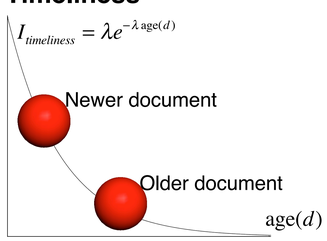
Topic Relevancy



$$I_{topic} = \Pr(\text{interesting}(d) \mid \text{keywords}(d))$$

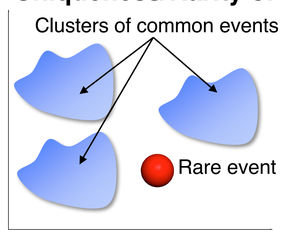
A document must be relevant to a topic of interest for it to be interesting.

Timeliness



Newer documents are generally more interesting than older documents.

Uniqueness/Rarity of Event



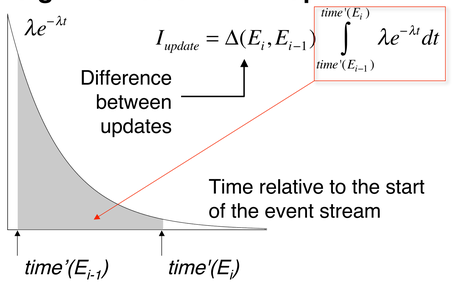
A document describing a rare or unique event is interesting.

Source Reputation

$$I_{reputation}(source) = \sum_{d' \in \text{Documents from } source} \lambda e^{-\lambda t - time(d')} I(d')$$ ← Interestingness of document d'

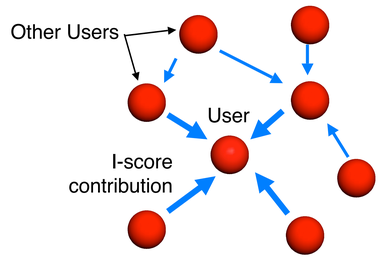
Data sources (e.g., authors) that have produced interesting documents in the past are likely to produce interesting articles in the future. Older articles have less significance than newer articles. Additionally, authors that are often cited by other authors of authority have been indirectly identified by their community to have produced interesting articles.

Significance of Event Updates



A document providing significant new information to an event last updated a long time ago is interesting. Additionally, documents that first describe a new event are interesting.

Collaborative Scoring



A document classified as interesting to one user should be also be interesting to other users with similar interests. Additionally, documents that are similar to each other should have similar interestingness scores for the same user. Social network analysis can help address this feature.

Threat Assessment

$$I_{threat} = \frac{2}{1 + e^{-\lambda Threat(E)}} - 1$$

$$Threat(E) = P(E \mid d) \times \text{Impact}(E)$$

Bayesian Belief Networks

Event Database

In the case of documents describing disaster events, such as disease outbreaks, those that describe highly probable events with significant impact are interesting.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.